Available online at www.sciencedirect.com

ScienceDirect

Journal of Biomedical Informatics 41 (2008) 1088–1100

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Word sense disambiguation across two domains: Biomedical literature and clinical notes

Guergana K. Savova [a,*], Anni R. Coden [b], Igor L. Sominsky [b], Rie Johnson [c,1], Philip V. Ogren [a], Piet C. de Groen [a], Christopher G. Chute [a]

[a] *Division of Biomedical Informatics, Mayo Clinic College of Medicine, 150 Third Street SW, Rochester, MN 55902, USA*
[b] *IBM, T.J. Watson Research Center, Hawthorne, New York, USA*
[c] *RJ Research Consulting, Tarrytown, New York, USA*

Received 26 December 2007
Available online 4 March 2008

Abstract

The aim of this study is to explore the word sense disambiguation (WSD) problem across two biomedical domains—biomedical literature and clinical notes. A supervised machine learning technique was used for the WSD task. One of the challenges addressed is the creation of a suitable clinical corpus with manual sense annotations. This corpus in conjunction with the WSD set from the National Library of Medicine provided the basis for the evaluation of our method across multiple domains and for the comparison of our results to published ones. Noteworthy is that only 20% of the most relevant ambiguous terms within a domain overlap between the two domains, having more senses associated with them in the clinical space than in the biomedical literature space. Experimentation with 28 different feature sets rendered a system achieving an average *F*-score of 0.82 on the clinical data and 0.86 on the biomedical literature.

Keywords: Natural language processing; Word sense disambiguation; Information extraction; Biomedical natural language processing; Artificial intelligence; Machine learning

1. Introduction

The amount of information generated in the biomedical sciences and medical practice has expanded exponentially in recent years. Of particular interest to us is the ever increasing amount of textual clinical data found in electronic patient medical records. At the Mayo Clinic the number of electronic clinical notes has increased from none in 1990 to approximately 25 million currently. Accessing and retrieving information from this growing repository of largely unstructured information has become both more challenging and more important. Similarly, retrieval of information from the biomedical literature has become increasingly important as the value of that literature grows. As Shatkay and Feldman [1] point out:

"Almost every known or postulated piece of information pertaining to genes, proteins, and their role in biological processes is reported somewhere in the vast amount of published biomedical literature... Moreover, automated literature mining offers a yet untapped opportunity to integrate many fragments of information gathered by researchers from multiple fields of expertise into a complete picture exposing the interrelated roles of various genes, proteins, and chemical reactions in cells and organisms." (p. 822).

The biomedical community has begun processing the large amounts of unstructured textual data for a number of tasks ranging from indexing to mining for novel information. The unstructured textual data resides in *scholarly scientific articles* as well as in *clinical record repositories*, e.g. clinical notes, pathology notes, radiology notes, treatment notes, clinical trials records. Mining and linking

* Corresponding author. Fax: +1 507 284 0360.
E-mail address: savova.guergana@mayo.edu (G.K. Savova).
[1] This work was done when the fourth author, Rie Johnson, was at IBM T.J. Watson Research Center.

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.02.003

information from both has the potential to greatly advance the scientific discovery process and transform conventional clinical care into personalized treatment plans. Clinical repositories are often the only source of information for evidence-based medicine, critical for devising personalized treatment plans as exemplified by issues in connection with devising dosage plans for the drug Warfarin.[2]

A number of investigative efforts focus on Natural Language Processing (NLP) techniques for processing biomedical scholarly literature, for example those stored on PubMed (Entrez PubMed[3]; [2]). Krallinger and Valencia [3] and Shatkay and Feldman [1] provide an extensive overview of the available life sciences search tools and biomedical NLP components along with compelling motivation for text mining in the biomedical domain. Two examples of NLP tools are MetaMap developed at the National Library of Medicine (NLM) [4] and ABNER.[4]

Despite the advancements in automated searches for the biomedical literature, progress with processing clinical data has been relatively slow. There are only a small number of research teams working on NLP approaches for that domain. This is largely due to confidentiality provisions for patient data. An example of an NLP system for the clinical domain is the Medical Language Extraction and Encoding System (MedLEE) which processes radiology mammography reports and discharge summaries at the Columbia-Presbyterian Medical Center [5]. Another example is the Cancer Text Information Extraction System[5] (caTIES) built at the University of Pittsburg as part of the Cancer Biomedical Informatics Grid[6] (caBIG) project at the National Cancer Institute. caTIES encodes information from free text surgical pathology reports to populate caBIG-compliant data structures.

In general, NLP techniques utilize ontologies as their knowledge bases for semantic processing. Semantic processing goes beyond simple string identification as it discovers the unique ontology entry to which the string belongs. In biomedicine, the Unified Medical Language System (UMLS) [6], or a subset of it, is used frequently as the ontology of choice. A source of errors in any NLP-based search engine, including systems for the biomedical domain, is the mapping of a word token or phrase to multiple concepts within an ontology [3,7–10]. For example, the mention "ms" maps to the following 2007AC UMLS concepts: *Marinesco-Sjogren syndrome, metric system, Mississippi (geographic location), mitral valve stenosis, Montserrat, multiple sclerosis, milliseconds, MTR gene, academic degree (Master of Science), supernumerary mandibular left primary canine, microbiology susceptibility domain, multiple sclerosis (susceptibility to)*. The unique mapping of a mention to a single concept that belongs to a set of concepts is known as word sense disambiguation (WSD). Words that can potentially be mapped to multiple concepts within an ontology or other sense inventories are referred to as ambiguities.

Ambiguity in the biomedical domain is well documented. Weeber and colleagues [7] report that there are 7400 ambiguous concepts in the UMLS. Xu and colleagues [10] show empirical results that ambiguity among gene names, English words and other biomedical terms is as high as 99.8%. Sehgal and colleagues [11] created a list of 1051 human gene terms that overlap with generic English meanings.

In this paper, we investigate the problem of WSD in the biomedical domain. Like in other domains, words and phrases are disambiguated by their context, i.e. patterns of words or concepts surrounding the word or phrase with an ambiguous sense. Examples of patterns are terms adjacent to the ambiguous word, their part-of-speech tags or an associated semantic class. These patterns are referred to as features. Our goals for the current study are: (1) to identify the top 50 relevant ambiguities in a large corpus of free text clinical notes and compare the results to those of a dataset representing the biomedical literature domain; (2) to apply and evaluate a state-of-the-art WSD machine learning algorithm on small datasets (up to 120 sense-labeled instances for each ambiguity) from both domains; (3) to identify the most productive features—patterns within text surrounding terms with an ambiguous sense—for this technique across domains and ambiguities; and (4) to compare our WSD results against results published elsewhere.

The paper is organized as follows. The Related Work section sets the background for the WSD-related research in general. Section 3 describes the data sets used in our work, the algorithm, the relevant features considered as input to the algorithm, the evaluation procedure and the metrics. Section 4 reports on the outcomes from our experiments. Section 5 offers our insights based on the results. We conclude with a summary of our work presented in this paper.

2. Related work

WSD investigation in the general domain has a long and rich history. Agirre and Edmonds [12] provide an excellent overview of this research since its early days in the late 1940s. The techniques applied to solve the WSD problem fall broadly into three categories: methods using knowledge bases and rules, methods using machine learning (unsupervised and supervised) and methods combining both. *Knowledge-based approaches* rely on the use of external lexical resources in the form of knowledge bases, dictionaries and/or thesauri. Such approaches typically incorporate some semantic similarity metric between a word and a sense derived from the lexical resources. *Supervised learning methods* need label-annotated information to build the models. The input to the algorithm is a vector of features usually extracted from the surrounding context.

[2] http://www.bio-itworld.com/issues/2007/sept/first-base.
[3] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed.
[4] http://pages.cs.wisc.edu/~bsettles/abner/.
[5] https://cabig.nci.nih.gov/tools/caties.
[6] http://SPECIALIST.nlm.nih.gov/.

The output is a sense assignment from a predetermined set of labels. *Unsupervised learning methods*, on the other hand, work with unlabeled data mainly through clustering techniques which aim at forming groups of related instances. Schutze [13] refers to this task as word sense discrimination to differentiate it from word sense disambiguation which includes the additional step of sense labeling. The input to the algorithm is a vector of features derived from the surrounding context. The output, however, is a set of clusters that do not link to a label from any predetermined set. For any machine learning approach, feature selection is critical for its performance. A variety of features have been experimented with—tokens from the surrounding context within a given window size, collocations, *n*-grams (sequences of *n* words), part-of-speech tags, orientation, distance, semantic information. *Combination, or hybrid, methods* explore the strength of knowledge-based and machine learning methods in a complementary fashion. For references for each method, see [12]. WEKA [14] is a Java package that implements a variety of machine learning techniques.

Schuemie, Kors, and Mons [15] review the applicability of general WSD algorithms to the biomedical domain and list the available sense-tagged datasets in that domain. The biomedical WSD methods are similar to these in the general domain with the exception of the lexical knowledge bases used. UMLS [6] is the predominant choice as a sense inventory in the biomedical domain. The National Library of Medicine (NLM) has developed a publicly available WSD test set [7] which has been used extensively in the biomedical WSD research community. The test set contains 50 ambiguities with 100 instances each manually sense-tagged against the UMLS. We describe this set in greater detail in Section 3 of this paper.

Here, we highlight some of the recent WSD research in the biomedical domain which is mainly in the non-clinical arena. Our focus stays on supervised machine learning methods as applied on small data sets (up to 120 sense-labeled instances for each ambiguity), an effort similar to ours. For an extensive list of references for the biomedical WSD, consult [15]. Xu and colleagues [10] use Support Vector Machines (SVM) for four ambiguous biomedical abbreviations to study the effect of sample size (20, 40, 80, and 120 sense-labeled training instances), sense distribution and degree of difficulty on the performance of the classifier. SVMs use functions to transform non-linear space into linear and construct a maximum margin hyperplane that provides the greatest separation between the classes. Xu and colleagues conclude that "(1) increasing the sample size generally reduces the error rate, but this was limited mainly to well-separated senses, (2) the sense distribution did not have an effect on performance when the senses were separable, (3) when there was a majority sense of over 90%, the WSD classifier was not better than use of simple majority sense, (4) error rates were proportional to the similarity of senses, and (5) there was no statistical difference between results when using a 5-fold or 10-fold cross-validation method."

Liu et al. [8] experimented with Naïve Bayes, decision lists, their adaptation of decision lists and their version of mixed supervised learning. Naïve Bayes classifiers pick the category with the highest probability and work off the assumption that the features are independent of one another. Decision lists are a sequence of tests applied to each incoming vector until it matches a condition or the end of the list is reached. The features that they use are co-occurring words, orientation, distance, collocations and varied window sizes. They experimented with 22 terms from the NLM WSD test set and concluded that to achieve good results with supervised WSD, there need to be at least a few dozen instances for each sense. Among the features, collocations were found to be the most productive. In general, they infer that in the biomedical literature domain the window size could be expanded to the entire paragraph, which is in sharp contrast to the general English domain.

Leroy and Rindflesch [9] focus on using UMLS symbolic knowledge to build a WSD classifier from small datasets. They trained a Naïve Bayes model for 15 words from the NLM WSD set. The inclusion of the UMLS semantic type as a feature yielded mixed results—from 8% deterioration to 29% improvement over the most frequent (a.k.a. majority) sense baseline. Other features they used in the study are part-of-speech tags, phrase heads, semantic relations between the unambiguous semantic types and its derivative sense activation. In a follow-up study, Leroy and Rindflesch [16] experimented with additional algorithms (decision tree and neural networks) but the results did not differ significantly.

A recent method using an external knowledge base was proposed by Humphrey and colleagues [17]. They statistically associate Journal Descriptor Indexing (JDI) assignments with words in a training set of Medline citations. Their experiments use 45 ambiguous words from the NLM WSD test set. The reported result of 0.7873 precision is a substantive improvement from their reported 0.2492 baseline.

Pakhomov and colleagues [18] focus on WSD in the clinical domain. They experiment with abbreviation and acronym disambiguation and apply a combination of supervised and unsupervised methods. They used the contexts harvested from the Internet to collect training data. These contexts were then applied to disambiguate the sense of the abbreviations in clinical notes.

Our research effort is similar to previous ones in that it uses machine learning techniques to build WSD classifiers from small datasets. What differentiates it from the other work is that we extend the technique to the unstructured textual data from the clinical domain. For that, we use an algorithm similar to SVMs which has shown state-of-the-art results for WSD in the general domain. We conduct extensive experimentation on clinical data to find the most productive features to train the classifier on. We also compare the performance of the algorithm on the NLM WSD

set to previously published results including exploration of the different usage of the same terms within a clinical setting versus a medical literature context.

3. Methods

3.1. Data sets

To achieve our study goals, two different datasets were considered—the NLM WSD test set used by other researchers, and a clinical test set developed by the Mayo Clinic. The datasets are described in detail later in this section. The total number of unique ambiguities we experimented with is 83—41 are unique to the Mayo Clinic dataset, 33 are unique to the NLM WSD dataset and 9 overlap between the two datasets. Each ambiguity is represented by 100 instances except two ("ms" and "sob" from the Mayo Clinic dataset) which have 1000 instances.

The NLM WSD test set is described in [7]. That set consists of 50 ambiguous words that occur in medical journal abstracts from Medline that have been manually tagged with UMLS concepts. Each ambiguity is represented by 100 abstracts. Table 1 summarizes the terms included in the NLM WSD set, the sense distributions and inter-annotator agreements (IAA). IAA is reported as kappa per the NLM. The evaluation against the NLM WSD test set provides the basis for a comparative study of different algorithms in the biomedical domain.

In our study, we excluded NLM WSD dataset terms with majority sense greater than 97% as these cases are considered extreme sense distributions. The excluded eight terms are *association, energy, fluid, inhibition, secretion, single, surgery* and *transient*. Some of them, e.g. *association* and *fluid*, present an interesting case of where all instances belong to one sense yet IAAs are low. For a full description of the NLM WSD, consult [7]. Similar subsets of the NLM WSD test set were used in [8] and [9]. Liu et al. [8] motivate the exclusion of NLM WSD set terms from their study as follows: "we excluded 12 [terms] that Weeber et al. [7] considered problematic as well as 16 terms in which the majority sense occurred with over 90% of instances". Leroy and Rindflesch [9] "selected 15 words from the NLM dataset for which the most frequent sense was correct in less than 65% of the instances".

Additionally, we developed a corpus of 50 ambiguities derived from Mayo Clinic clinical notes. We followed the same methodology as the one used to create the NLM WSD set. The sense inventory was UMLS. First, we identified the top 1000 most frequent ambiguities in the corpus of Mayo Clinic clinical notes from the year of 2002. Frequency was the count of the number of occurrences of each term in the 2002 clinical notes. We used MetaMap to map all terms from the notes to the UMLS. Terms with multiple mappings were considered ambiguous. After ranking them by frequency, the list of the top 1000 most frequent ambiguities was compiled. That list was then submitted to medical index retrieval experts to rank for relevancy for retrieval purposes. The top 50 ambiguities judged relevant by the experts were included in the final set. 48 ambiguities have 100 instances each and 2 ("ms" and "sob") have 1000 instances each. All instances were randomly picked from the clinical notes. The subsequent sense annotation was done manually by four experts against UMLS. A "none of the above" category was added for those instances that did not have any sense representation in the UMLS to follow the NLM methodology for sense-tagged test set creation [7]. Consensus discussions determined the final sense assignment for the instances for which the experts disagreed thus creating the final consensus set that was used in our experiments. Table 2 shows the terms included in the Mayo WSD set, the sense distributions and inter-annotator agreements (IAA). IAA is computed as kappa [19,20].

The kappa values are weighted as they take into account the actual distribution of the senses for the computation of the expected agreement $(P(E))$. For a detailed description and examples see [21]. Kappa values are computed as

$$k = \frac{(P(A) - P(E))}{(1 - P(E))} \tag{1}$$

where

$$P(A) = \frac{\text{number of agreed instances}}{N} \tag{2}$$

$$P(E) = \sum_{j=1\ldots m} P_j^2 \tag{3}$$

where

$$P_j = \frac{C_j}{Nk} \tag{4}$$

- m is the number of senses for a given term
- Cj is the number of times an instance is assigned j sense
- N is the total number of instances for all senses
- k is the number of annotators

In our case, k is 2, and N is 100 except for *ms* and *sob* where N is 1000.

Of note is that one could compute $P(E)$ by assigning equal probability to each sense (non-weighted kappa). Non-weighted kappa values are used in some studies [22] but we feel the weighted kappa provides a better estimate of the actual inter-annotator agreement as it takes into account the actual sense frequencies. The average kappa value for the Mayo WSD set is $k = 0.54$ and is similar to the average kappa value of the NLM WSD set ($k = 0.47$). Note that in our experiments we used the final consensus set as described above.

For each dataset, we created one additional version in which we removed senses with fewer than three training instances (referred to as pruned sets). This excludes underrepresented senses as they present a challenge for any algorithm to learn. It also allows for a direct comparison with results reported elsewhere as this same technique was used by others [8].

Table 1
Ambiguous terms included in the NLM WSD dataset with sense distribution and inter-annotator agreement (IAA)

Ambiguity	Number of instances	Number of senses	Sense1 (majority sense, in %)	Sense2 (in %)	Sense3 (in %)	Sense4 (in %)	Sense5 (in %)	None (included in sense1–sense5) (in %)	IAA (kappa)
Adjustment	100	4	62	18	13	7		7	0.43
Association	100	1	100					100	0.26
Blood pressure	100	3	54	44	2				0.11
Cold	100	5	86	6	5	2	1	5	0.48
Condition	100	3	90	8	2			8	0.06
Culture	100	2	89	11					0.78
Degree	100	3	63	35	2			35	0.62
Depression	100	2	85	15				15	0.80
Determination	100	2	79	21				21	0.30
Discharge	100	3	74	25	1			25	0.80
Energy	100	2	99	1					0.17
Evaluation	100	2	50	50					0.32
Extraction	100	3	82	13	5			13	0.27
Failure	100	3	71	25	4			71	0.23
Fat	100	3	71	27	2			27	0.26
Fit	100	2	82	18				82	0.86
Fluid	100	1	100						0.07
Frequency	100	2	94	6				6	0.08
Ganglion	100	2	93	7					0.54
Glucose	100	2	91	9					0.21
Growth	100	2	63	37					0.41
Immunosuppression	100	2	59	41					0.58
Implantation	100	3	81	17	2			2	0.71
Inhibition	100	3	98	1	1			1	0.11
Japanese	100	3	73	21	6			21	0.54
Lead	100	3	71	27	2			71	0.84
Man	100	4	58	33	8	1		8	0.53
Mole	100	3	83	16	1			16	0.91
Mosaic	100	3	52	45	3			3	0.20
Nutrition	100	4	45	28	16	11		11	0.32
Pathology	100	3	85	14	1			1	0.67
Pressure	100	2	96	4				4	0.30
Radiation	100	3	61	37	2			2	0.49
Reduction	100	3	89	9	2			89	0.38
Repair	100	3	52	32	16			32	0.79
Resistance	100	2	97	3				97	0.70
Scale	100	2	65	35				35	0.57
Secretion	100	2	99	1					0.07
Sensitivity	100	4	49	49	1	1		49	0.53
Sex	100	3	80	15	5				0.78
Single	100	2	99	1					0.71
Strains	100	3	92	7	1			7	0.49
Support	100	3	90	8	2			90	0.44
Surgery	100	2	98	2					0.26
Transient	100	2	99	1					0.23
Transport	100	3	93	6	1			6	0.72
Ultrasound	100	2	84	16					0.48
Variation	100	2	80	20					0.29
Weight	100	3	47	29	24			47	0.73
White	100	3	49	41	10			10	0.86
Average		2.64	78.04						0.47

None (included in sense 1–sense 6) indicates the percentage of instances that belonged to the none of the above sense category, which is one of the six sense categories.

3.2. Algorithm and features

In modern statistical machine learning, it has become standard to train predictors based on the empirical risk minimization principle, i.e., to obtain a predictor by minimizing prediction error (called empirical risk) measured by an appropriate loss function on the labeled training examples with appropriate regularization. This family of classifiers includes state-of-the-art methods such as SVMs. In our experiments, we used a classifier that belongs to the same family and employs a modification of Huber's loss as the loss function and stochastic gradient descent as the

Table 2
Ambiguous terms included in the Mayo WSD dataset with sense distribution and inter-annotator agreement (IAA)

Ambiguity	Number of instances	Number of senses	Sense1 (majority sense, in%)	Sense2 (in %)	Sense3 (in %)	Sense4 (in %)	Sense5 (in %)	Sense6 (in %)	Sense7 (in %)	Sense8 (in %)	Sense9 (in %)	Sense10 (in %)	Sense11 (in %)	Sense12 (in %)	None (included in sense1–sense12, in %)	IAA (weighted kappa)
ac	100	12	59	9	7	5	5	4	3	2	2	2	1	1	0	0.35
Adjustment	100	5	82	12	4	1	1								1	0.35
Affect	100	2	50	50											0	0.88
Aid	100	5	47	39	9	4	1								9	0.26
Ape	100	6	84	6	3	3	3	1							3	0.69
Aspiration	100	2	59	41											0	0.98
Block	100	9	34	19	13	13	12	3	3	2	1				12	0.01
Burn	100	6	82	8	7	1	1	1							1	1.00
Cat	100	3	50	48	2										0	0.94
Cervical	100	4	43	41	14	2									0	0.83
cf	100	B	62	17	9	4	4	2	1	1					1	0.60
Cold	100	6	64	14	14	4	3	1							1	0.85
Compression	100	4	56	38	4	2									0	0.85
Condition	100	5	84	9	5	1	1								0	0.37
Dilatation	100	2	52	48											0	0.94
Discharge	100	2	65	35											0	0.94
Drain	100	4	47	29	23	1									0	0.72
Dress	100	6	35	24	23	11	6	1							0	0.27
Drink	100	4	54	42	3	1									0	0.12
Fast	100	4	85	13	1	1									0	0.89
Fistula	100	3	78	20	2										0	0.33
Fit	100	5	68	13	11	6	2								6	0.70
Glass	100	4	55	38	5	2									0	0.20
Grade	100	6	38	31	17	9	4	1							0	0.36
Interaction	100	5	72	20	4	3	1								4	0.78
Iron	100	4	70	24	3	3									0	0.76
Irritate	100	3	84	15	1										0	0.34
iv	100	4	45	23	20	12									0	0.72
Lead	100	5	47	26	17	7	3								7	0.47
Lift	100	7	87	5	2	2	2	1	1						0	0.10
ms	1000	10	93.1	4.7	0.6	0.5	0.3	0.2	0.2	0.2	0.1	0.1			0.2	0.96
pa	100	7	56	34	4	2	2	1	1						2	0.82
Pack	100	5	79	15	3	2	1								3	−0.29
Patch	100	B	73	18	5	2	1	1							0	0.42
Plaque	100	3	49	39	12										0	0.26
ra	100	9	40	23	20	6	5	3	1	1	1				5	0.61
Relative	100	2	53	47											0	0.42
Sense	100	4	43	26	18	13									0	0.32
Sensitivity	100	7	39	22	21	12	3	2	1						3	0.34
Sob	1000	2	98.3	1.7											0	1.00
Splint	100	4	70	28	1	1									1	0.88
Spot	100	3	39	20	14	11	7	5	3	1					20	0.60
Stage	100	10	28	17	16	12	7	6	5	4	3	2			5	0.10
Strain	100	9	35	18	17	10	7	6	5	1	1				18	0.44
Stress	100	3	77	22	1										22	0.94
Support	100	B	38	25	13	11	9	4							11	0.46
Tear	100	7	59	17	11	6	3	3	1						3	0.84
Transfer	100	2	84	16											0	−0.09
Valve	100	3	52	43	5										0	1.00
Vesicle	100	3	73	25	2										0	−0.54
Average		5.1	60.3													0.54

None (included in sense1–sense12) indicates the percentage of instances that belonged to the none of the above sense category, which is one of the 12 sense categories.

training algorithm. Ando [23] showed that this configuration (referred to as "supervised baseline" there) achieves state-of-the-art performance on WSD tasks in the general domain. Alternatively, one could use SVM, which is also known to achieve state-of-the-art performance on WSD tasks [24]. Zhang [25] should be consulted for the technical detail of our configuration.

We experimented with a number of features and combinations between them. An example of an ambiguity and extracted features is presented in Fig. 1. A more challenging example for the algorithm is "*I reviewed with Ms. Smith her knowledge of MS.*" where there are two target ambiguities—*Ms.* and *MS*. The feature vectors for the targets share context although the senses are different. Cases like these motivate the usage of finer grained features, e.g. orientation and distance; however, they introduce potential vector sparsity.

- *Bag of Words (BOW)* is the representation of the context by the unique words in it. The vector thus created is a non-weighted feature representation as the frequency of the features, i.e. words, is not taken into account.
- *Bag of Stemmed Words (BSW)* lists the unique stems within the context window. It is the stemmed version of BOW. Stemming was performed using NLM's Lexical Variant Generator (LVG)[7] tool.
- *Part-of-speech (POS) tags*, e.g. noun, verb, adjective, were obtained using the POS tagger developed at IBM and trained on Mayo and Linguistic Data Consortium data [26].
- *Window size* is the number of tokens representing context surrounding the ambiguity. We experimented with a window of 5, 10, and 50 tokens on both sides of the target ambiguity.
- *Orientation* is the location of the feature in regard to the target ambiguity. Possible values are left or right.
- *Distance* is the proximity of the feature in regard to the target ambiguity.
- Medical Subject Headings (MeSH)[8] semantic classes were assigned to named entities. Within MeSH, descriptors are organized in 16 categories.[9] A dictionary lookup tool[10] was used to make the assignments between the text and the MeSH terminology. Note, that none of the pre-existing MeSH labels were used. Only words with unambiguous assignments of a MeSH descriptor were used. The semantic class assigned to a word, is the label of the category to which its descriptor belongs. In the example in Fig. 1, "Diazepam" maps to D03.438.079.080.070.216 with a semantic class of D03 which is one of the 16 MeSH categories. That semantic class was used as the feature.
- *Named Entities* in our study are defined as the categories of disorders, drugs, findings and procedures. For the named-entity recognition, we used Mayo's named-entity recognizers built jointly with IBM. It is a classifier for the four categories. However, we have not formally evaluated its performance yet.
- *Metadata* represents the section heading and the medical specialty of the clinical note the ambiguity occurs in. This feature was not included in the experiments with the NLM WSD set as it is unique to the clinical notes.

We additionally removed stopwords (non-content words like *is, a, an*), punctuation and low frequency words from the features. To remove stopwords and punctuation, we created a standard list of prepositions, determiners and punctuation symbols. In order to remove non-discriminative words or non-discriminative stemmed words and to retain as features only the most discriminative words or stemmed words, we implemented a version of tf*idf aiming at discovering the most distinctive words for each sense:

$$tf = \frac{\text{word frequency}}{\text{total number of words}} \quad (5)$$

$$idf = \log 2\left(\frac{\text{number of total instances}}{\text{number of instances where the word appears}}\right) \quad (6)$$

*tf*idf filtering* is done per word. All words with the lowest tf*idf score were filtered out. Note that frequencies were computed within the windows of 50 for each instance.

Table 3 summarizes the 28 experiments we ran. Each of them had a different feature combination in order to explore which features were the most discriminating. Each combination measures the contribution of a particular feature. Note that not all possible combinations were executed as we felt feature contributions could be estimated based on a subset of all possible experiments:

- ◦ Window size: T1, T2, T3, T4, T9, T10, T11, T13, T14, T16, T17, T18, T19, T20, T21, T22, T23, T24, T25, T26
- ◦ Punctuation and stoplist words: T2 and T5
- ◦ tf*idf filter: T1, T2, T3, T4, T16, T17, T19
- ◦ Orientation: T2, T5, T6, T7
- ◦ Orientation and distance: T2, T5, T8, T9
- ◦ Stemming: T1, T4, T11, T16
- ◦ Metadata: T3, T13, T14, T21, T23, T24

3.3. Evaluation procedure and metrics

We evaluated the performance of our algorithm against a baseline evaluation method—the Frequency method. By default, the Frequency method assigns the most frequent sense to all instances of a target ambiguity. A system is considered to perform well if it outscores the baseline.

[7] http://SPECIALIST.nlm.nih.gov.
[8] http://www.nlm.nih.gov/mesh/meshhome.html.
[9] http://www.nlm.nih.gov/mesh/intro_trees2006.html.
[10] http://uima.lti.cs.cmu.edu:8080/UCR/ViewComponentAction.do?componentId=33.

	distance (window=10)	orientation wrt target ambiguity
She	10	left
states	9	left
she	8	left
becomes	7	left
very	6	left
tired	5	left
and	4	left
upset	3	left
with	2	left
her	1	left
MS	target ambiguity	
.	1	right
She	2	right
stated	3	right
that	4	right
the	5	right
tiredness	6	right
and	7	right
Diazepam	8	right
were	9	right
her	10	right
primary	outside of window	
issue	outside of window	
.	outside of window	
BOW	she|states|becomes|very|tired|and|upset|with|her| .|stated|that|the|tiredness|Diazepam|were|	
stopwords	she|and|with|her|that|the|were|	
BOW without stopwords and punctuation	states|becomes|very|tired|upset|stated|tiredness|Diazepam|	
BSW without stopwords and punctuation	state|become|very|tire|upset|Diazepam|	
POS tags (Penn Treebank) within the window	PRP|VBP|JJ|CC|IN|PRP$|VBD|DT|NN|	
Orientation	states_left|becomes_left|stated_right|tiredness_right	
Distance	states_9|becomes_7|stated_3|tiredness_6	
Named Entities	tired_C0557875|upset_C0700361|diazepam_C0012010	
Metadata	specialty_neurology|sectionHeading_hpi	
tf*idf for "tired" in the MS corpus	word frequency = 10; total number of words = 1000;	
	number of total instances = 10 senses for MS; number of instances where the word appears = 1	
	tf = 0.01; idf = 3.32; tf*idf = 0.33	
tf*idf for "stated" in the MS corpus	word frequency = 100; total number of words = 1000;	
	number of total instances = 10 senses for MS; number of instances where the word appears = 9	
	tf = 0.1; idf = 0.15; tf*idf = 0.015	
MeSH terms	Diazepam_D03 (note that the entire tree for Diazepam is D03.438.079.080.070.216)	

Fig. 1. Example of an ambiguity and its features within a window of 10. Context is "*She states she becomes very tired and upset with her MS. She stated that the tiredness and Diazepam were her primary issues.*" Target ambiguity is the underlined mention of MS.

The algorithm and feature sets were tested through a 10-fold cross-validation without replacement as described in [27]. The labeled set is divided into 10 parts, 8 of which are used for training and 2 for testing. This is repeated 10 times and the average performance from the 10 tests is recorded. Several metrics were computed during our evaluation process:

- TP—true positives, or the labels for a given sense correctly recognized by the algorithm
- AP—actual positives, or the total occurrences for a given sense in the test set
- PP—predicted positives, or the reported occurrences for a given sense made by the algorithm

$$\text{Precision} = \frac{\text{TP}}{\text{PP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{AP}} \tag{8}$$

$$F - \text{score} = \frac{(2 * \text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{9}$$

We computed the average of the *F*-score micro-averages of all the terms per feature set. The micro-average was computed by adding the TP, AP, and PP for all runs in that category and then using these sums in the *F*-score calculations. The average *F*-score was calculated from the 10-fold evaluation results. 95% confidence intervals are reported for

Table 3
Feature combinations and experiment numbering

Experiment/feature	T1	T2	T3	T4	T5	IB	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24	T25	T26	T27	T28
Bag of words	X	X	X	X	X	X	X	X	X	X											X	X					X	
Bag of stemmed words											X					X	X	X	X	X						X		X
MeSH classes													X	X	X								X	X	X			
MeSH classes mapped to category																											X	X
POS												X			X													
Metadata (section heading and service code)														X							X	X		X			X	X
Named entities													X	X	X								X	X	X			X
Stop words and punctuation removed	X	X	X	X		X		X			X					X	X	X	X	X	X	X				X	X	X
tf*idf filtered			X	X												X		X	X	X	X	X				X	X	X
orientation						X	X	X	X	X		X			X				X	X								
Distance								X	X	X		X			X													
Window size 5	X		X								X										X		X	X		X	X	
Window size 10		X		X	X	X	X	X	X							X	X		X			X			X			X
Window size 50										X		X	X	X	X			X		X								

the evaluation metrics computed using the so-called "exact" confidence intervals [28].[11] The tests for statistical significance reported in this paper use a *t*-test for paired two sample means and a level of significance of 0.05. The null hypothesis is that there is no difference.[12]

4. Results

Results are reported on the two datasets described in Section 3—the NLM WSD set and the Mayo WSD set. Results on the former provide the basis for cross-study comparisons. Results on the latter provide evidence for the transferability of the algorithm across domains along with most productive feature sets.

Our first goal was to identify the top 50 ambiguities in the clinical domain and compare them to these in the biomedical literature domain. The average number of senses for the NLM WSD dataset is 2.64, while that for the Mayo Clinic dataset is 5.1. Majority sense for the NLM WSD dataset is 78.04%, while that for the Mayo Clinic WSD dataset—60.3%. There are only 9 overlapping terms between the NLM and the Mayo Clinic WSD dataset—*adjustment, cold, condition, discharge, fit, lead, sensitivity, strain,* and *support*—pointing to domain characteristics. Additionally, the differences across the two domains pose different challenges—each sense from the Mayo WSD dataset is expected to have fewer instances spread among the 100 sense-tagged instances resulting in smaller amounts of training data.

We ran our WSD algorithm on the NLM dataset (biomedical literature domain) with the features and their combinations as described in Table 3 excluding the Metadata (Section heading and service code) feature as it is not relevant to biomedical literature. We then compared our best results to these published elsewhere. Table 4 displays the summary results on the NLM WSD set from the current study and the ones reported in [8] and [9]. Column 1 lists the ambiguity, column 2 is the majority sense for that ambiguity, column 3 is the inter-annotator agreement (IAA) as a kappa value, followed by the *F*-scores from each study along with 95% confidence intervals.

There are 28 ambiguous terms that overlap among the Liu, Teller, and Friedman study [8], the Leroy and Rindflesch study [9] and our investigation. In nine cases, the best score is the one described in Liu, Teller, and Friedman study. In 19 cases, the best result is the one from the current experiments on the NLM WSD set.

Another goal in this study was to identify most productive features and their dependencies on the WSD algorithm, the ambiguous word itself and the ambiguous word as it is mentioned within a particular domain. The first set of results is in the biomedical literature domain. It turns out, as shown in Table 4, that the best *F*-scores were achieved with different feature sets not only across the three studies/algorithms—Liu et al. [8], Leroy and Rindflesch [9] and ours—but also within the studies themselves. Liu et al. [8] use a varied window size (2–4, 6, 8, 10) with several sets of features—words with oriented distance within the window, words with orientation within the window, words within the window, tree collocations, and oriented words within a window. Leroy and Rindflesch [9] use a combination of features—*heads of phrases, pos tags, the UMLS semantic type* of words in the same phrase as the ambiguous word, the *UMLS semantic type* of all other unambiguous words in the sentences and the *UMLS relations* of the ambiguous word with the other words. In our current study, the best results were achieved with a combination of a simple set of features—*bag of words, bag of stemmed words, tf*idf filters, stopwords removal, varied window size*. Although some

[11] http://www.graphpad.com/quickcalcs/.
[12] Microsoft Excel, *t*-test: paired two sample for means.

Table 4
Comparison results of best *F*-scores on NLM WSD set (IAA, inter-annotator agreement; NIS, not included in study; CI, 95% confidence interval)

Term	Majority sense (in %)	IAA (kappa)	Best *F*-score (entire NLM WSD set)	CI (entire NLM WSD set)	Best *F*-score (NLM WSD-pruned set)	CI (NLM WSD-pruned set)	Best *F*-score (Leroy and Rindflesch)	CI (Leroy and Rindflesch)	Best *F*-score (Liu, Tetter, and Friedman)	CI (Liu, Teller and Friedman)
Adjustment	62	0.43	0.75	0.65–0.83	0.75	0.65–0.83	0.62	0.52–0.71	NIS	NIS
Blood pressure	54	0.11	0.62	0.52–0.71	0.64	0.54–0.73	0.56	0.46–0.66	NIS	NIS
Cold	86	0.48	0.89	0.81–0.94	0.92	0.84–0.96	NIS	NIS	0.91	0.82–0.95
Condition	90	0.06	0.91	0.84–0.96	0.92	0.84–0.96	NIS	NIS	NIS	NIS
Culture	89	0.78	0.94	0.87–0.98	0.93	0.86–0.97	NIS	NIS	NIS	NIS
Degree	63	0.62	0.96	0.90–0.99	0.98	0.93–1.00	0.68	0.58–0.77	0.98	0.93–1.00
Depression	85	0.80	0.90	0.82–0.95	0.91	0.84–0.96	NIS	NIS	0.89	0.80–0.94
Determination	79	0.30	0.87	0.79–0.93	0.86	0.77–0.92	NIS	NIS	NIS	NIS
Discharge	74	0.80	0.95	0.89–0.98	0.96	0.90–0.99	NIS	NIS	0.91	0.82–0.95
Evaluation	50	0.32	0.77	0.67–0.85	0.77	0.67–0.85	0.57	0.47–0.67	NIS	NIS
Extraction	82	0.27	0.87	0.79–0.93	0.87	0.79–0.93	NIS	NIS	0.90	0.81–0.94
Failure	71	0.23	0.75	0.65–0.83	0.73	0.63–0.81	NIS	NIS	NIS	NIS
Fat	71	0.26	0.83	0.74–0.89	0.86	0.76–0.91	NIS	NIS	0.86	0.76–0.91
Fit	82	0.86	0.88	0.80–0.94	0.88	0.80–0.94	NIS	NIS	NIS	NIS
Frequency	94	0.08	0.96	0.90–0.99	0.96	0.90–0.99	NIS	NIS	NIS	NIS
Ganglion	93	0.54	0.96	0.90–0.99	0.96	0.90–0.99	NIS	NIS	NIS	NIS
Glucose	91	0.21	0.91	0.84–0.96	0.92	0.85–0.96	NIS	NIS	NIS	NIS
Growth	63	0.41	0.72	0.62–0.80	0.73	0.63–0.81	0.63	0.53–0.72	0.72	0.62–0.80
Immunosuppression	59	0.58	0.84	0.75–0.90	0.84	0.75–0.90	0.67	0.57–0.76	NIS	NIS
Implantation	81	0.71	0.96	0.90–0.99	0.96	0.88–0.98	NIS	NIS	0.90	0.82–0.95
Japanese	73	0.54	0.77	0.67–0.85	0.79	0.70–0.86	NIS	NIS	0.80	0.70–0.86
Lead	71	0.84	0.93	0.86–0.97	0.95	0.87–0.98	NIS	NIS	0.91	0.84–0.96
Man	58	0.53	0.87	0.79–0.93	0.87	0.78–0.92	0.80	0.71–0.87	0.91	0.84–0.96
Mole	83	0.91	0.96	0.90–0.99	0.97	0.91–0.99	NIS	NIS	0.91	0.84–0.96
Mosaic	52	0.20	0.87	0.79–0.93	0.85	0.76–0.91	0.69	0.59–0.78	0.88	0.79–0.93
Nutrition	45	0.32	0.49	0.39–0.59	0.45	0.35–0.55	0.53	0.43–0.63	0.58	0.48–0.68
Pathology	85	0.67	0.87	0.79–0.93	0.89	0.80–0.94	NIS	NIS	0.88	0.80–0.94
Pressure	96	0.30	0.96	0.90–0.99	0.96	0.90–0.99	NIS	NIS	NIS	NIS
Radiation	61	0.49	0.82	0.73–0.89	0.83	0.73–0.89	0.72	0.62–0.80	NIS	NIS
Reduction	89	0.38	0.91	0.84–0.96	0.92	0.84–0.96	NIS	NIS	0.91	0.84–0.96
Repair	52	0.79	0.89	0.81–0.94	0.91	0.84–0.96	0.81	0.72–0.88	0.76	0.66–0.84
Resistance	97	0.70	0.97	0.91–0.99	0.97	0.91–0.99	NIS	NIS	NIS	NIS
Scale	65	0.57	0.82	0.73–0.89	0.84	0.75–0.90	0.84	0.75–0.90	0.91	0.82–0.95
Sensitivity	49	0.53	0.92	0.85–0.96	0.91	0.82–0.95	0.70	0.60–0.79	NIS	NIS
Sex	80	0.78	0.91	0.84–0.96	0.90	0.82–0.95	NIS	NIS	0.90	0.81–0.94
Strains	92	0.49	0.93	0.86–0.97	0.94	0.86–0.97	NIS	NIS	NIS	NIS
Support	90	0.44	0.90	0.82–0.95	0.93	0.85–0.96	NIS	NIS	NIS	NIS
Transport	93	0.72	0.94	0.87–0.98	0.96	0.90–0.99	NIS	NIS	NIS	NIS
Ultrasound	84	0.48	0.88	0.80–0.94	0.90	0.82–0.95	NIS	NIS	0.88	0.79–0.93
Variation	80	0.29	0.89	0.81–0.94	0.89	0.81–0.94	NIS	NIS	NIS	NIS
Weight	47	0.73	0.78	0.69–0.86	0.80	0.71–0.87	0.71	0.61–0.80	0.78	0.69–0.86
White	49	0.86	0.73	0.63–0.81	0.75	0.65–0.83	0.62	0.52–0.71	0.76	0.65–0.83

combinations achieved best results for multiple ambiguities, there is no single feature set that performed best across all ambiguities.

Our next step was to evaluate the algorithm and the most productive feature sets in the clinical domain. Table 5 shows the results on the Mayo Clinic WSD dataset. Column 1 lists the ambiguity, column 2 is the majority sense, column 3 is the inter-annotator agreement as a kappa value, followed by *F*-score and 95% confidence intervals for the *F*-score.

Which are the most productive features in the clinical domain? The top 25 percentile are feature sets T18, T20, T02, T19, T17, T04, and T16 (Table 3 is to be consulted for the exact feature combinations). The features are *stemming, stopwords, and punctuation removed, tf*idf by stem or token, BOW* and *orientation*. The window sizes used were 5 and 10. None of the differences between the top 25 percentile *F*-scores are significant (p-values > 0.05).

The bottom 25 percentile are feature sets T24, T15, T25, T27, T23, and T13 (Table 3 is to be consulted for the exact

Table 5
Results for Mayo WSD dataset—best *F*-score, 95% confidence interval (CI) for *F*-score, feature set for best *F*-score (IAA, inter-annotator agreement)

Term	Majority sense (in %)	IAA (kappa)	Best *F*-score (entire corpus)	CI	Best *F*-score (Mayo WSD-pruned set)	CI
ac	59	0.35	0.80	0.71–0.87	0.84	0.75–0.90
Adjustment	82	0.35	0.93	0.86–0.97	0.93	0.86–0.97
Affect	50	0.88	0.97	0.91–0.99	0.97	0.91–0.99
Aid	47	0.26	0.76	0.66–0.84	0.75	0.65–0.83
Ape	84	0.69	0.88	0.80–0.94	0.89	0.81–0.94
Aspiration	59	0.98	0.92	0.85–0.96	0.94	0.87–0.98
Block	34	0.01	0.66	0.56–0.75	0.68	0.58–0.77
Burn	82	1.00	0.91	0.84–0.96	0.90	0.82–0.95
Cat	50	0.94	0.97	0.91–0.99	1.00	0.96–1.00
Cervical	43	0.83	0.88	0.80–0.94	0.89	0.81–0.94
cf	62	0.60	0.96	0.90–0.99	0.97	0.91–0.99
Cold	64	0.85	0.72	0.62–0.80	0.71	0.61–0.80
Compression	56	0.85	0.78	0.69–0.86	0.81	0.72–0.88
Condition	84	0.37	0.88	0.80–0.94	0.90	0.82–0.95
Dilatation	52	0.94	0.86	0.78–0.92	0.88	0.80–0.94
Discharge	65	0.94	0.92	0.85–0.96	0.95	0.89–0.98
Drain	47	0.72	0.71	0.61–0.80	0.73	0.63–0.81
Dress	35	0.27	0.70	0.60–0.79	0.73	0.63–0.81
Drink	54	0.12	0.74	0.64–0.82	0.76	0.66–0.84
Fast	85	0.89	0.90	0.82–0.95	0.93	0.86–0.97
Fistula	78	0.33	0.86	0.78–0.92	0.88	0.80–0.94
Fit	68	0.70	0.74	0.64–0.82	0.75	0.65–0.83
Glass	55	0.20	0.92	0.85–0.96	0.94	0.87–0.98
Grade	38	0.36	0.81	0.72–0.88	0.81	0.72–0.88
Interaction	72	0.78	0.86	0.78–0.92	0.87	0.79–0.93
Iron	70	0.76	0.88	0.80–0.94	0.89	0.81–0.94
Irritate	84	0.34	0.90	0.82–0.95	0.90	0.82–0.95
iv	45	0.72	0.65	0.55–0.74	0.65	0.55–0.74
Lead	47	0.47	0.67	0.57–0.76	0.68	0.58–0.77
Lift	87	0.10	0.93	0.86–0.97	0.99	0.94–1.00
ms	93.1	0.96	0.97	0.91–0.99	0.97	0.91–0.99
pa	56	0.82	0.89	0.81–0.94	0.96	0.90–0.99
Pack	79	−0.29	0.88	0.80–0.94	0.94	0.87–0.98
Patch	73	0.42	0.87	0.79–0.93	0.89	0.81–0.94
Plaque	49	0.26	0.97	0.91–0.99	0.97	0.91–0.99
ra	40	0.61	0.8S	0.80–0.94	0.89	0.81–0.94
Relative	53	0.42	0.95	0.89–0.98	0.96	0.90–0.99
Sense	43	0.32	0.82	0.73–0.89	0.84	0.75–0.90
Sensitivity	39	0.34	0.75	0.65–0.83	0.78	0.69–0.86
Sob	98.3	1.00	0.99	0.94–1.00	0.99	0.94–1.00
Splint	70	0.88	0.82	0.73–0.89	0.80	0.71–0.87
Spot	39	0.60	0.59	0.49–0.69	0.59	0.49–0.69
Stage	28	−0.10	0.53	0.43–0.63	0.53	0.43–0.63
Strain	35	0.44	0.59	0.49–0.69	0.56	0.46–0.66
Stress	77	0.94	0.95	0.89–0.98	0.96	0.90–0.99
Support	38	0.46	0.59	0.49–0.69	0.56	0.46–0.66
Tear	59	0.84	0.72	0.62–0.80	0.71	0.61–0.80
Transfer	84	−0.09	0.99	0.95–1.00	0.99	0.95–1.00
Valve	52	1.00	0.75	0.65–0.83	0.69	0.59–0.78
Vesicle	73	−0.54	0.79	0.70–0.86	0.80	0.71–0.87

Table 6
Overlapping ambiguities (NLM WSD dataset and Mayo Clinic WSD dataset): results with our algorithm

Term	Number of senses in NLM WSD set, majority sense, kappa	Number of senses in Mayo WSD set, majority sense, kappa	Best *F*-score (entire NLM WSD set)	Best *F*-score (NLM WSD-pruned set)	Best *F*-score (entire Mayo WSD set)	Best *F*-score (Mayo WSD-pruned set)
Adjustment	4 (62%; 0.43)	5 (82%; 0.35)	0.75	0.75	0.93	0.93
Cold	5 (86%; 0.48)	6 (64%; 0.85)	0.89	0.92	0.72	0.71
Condition	3 (90%; 0.06)	6 (84%; 0.37)	0.91	0.92	0.88	0.90
Discharge	3 (74%; 0.80)	2 (65%; 0.94)	0.95	0.96	0.92	0.95
Fit	2 (82%; 0.86)	5 (68%; 0.70)	0.88	0.88	0.74	0.75
Lead	3 (71%; 0.84)	5 (47%; 0.47)	0.93	0.95	0.67	0.68
Sensitivity	4 (49%; 0.53)	7 (39%; 0.34)	0.92	0.91	0.75	0.78
Strain	3 (92%; 0.49)	9 (35%; 0.44)	0.93	0.94	0.59	0.56
Support	3 (90%; 0.44)	6 (38%; 0.46)	0.90	0.93	0.59	0.56

feature combinations). The features are *MeSH classes and named entities, section heading, service code, part-of-speech tags, orientation*, and *distance* with varied window size (5, 10, and 50). T27 is the only feature set in that group that uses BOW in combination with a tf*idf filter by token. None of the differences between the bottom 25 percentile F-scores are significant (p-values > 0.05).

When we compare the performance of the algorithm on the entire corpus and the pruned version of the corpus, the results are, on the average, within one percentage point for both the NLM and Mayo data sets. This, in general, points to the stability of the contextual features as well as the robustness of the algorithm to outlier instances and in general. The performance of the algorithm on the pruned set falls within the confidence intervals for the performance on the entire set for the ambiguities with no instance occurrence of less than 3% (for example, *nutrition* from the NLM data set and *valve* from the Mayo data set).

Finally, Table 6 displays the summary results for the overlapping ambiguities between the NLM WSD dataset and the Mayo Clinic WSD dataset. This is our direct cross-domain comparison of our algorithm. The differences in the results tend to depend on the number of senses which implies number of training instances per sense. The fewer the senses, the more training instances there would be per sense. The most striking difference is the *F*-scores for *strain* and *support* for which the majority sense in the NLM dataset is much larger than the one in the Mayo Clinic dataset and there are fewer senses in the NLM dataset than in the Mayo Clinic dataset. In general, the overlapping ambiguities in the Mayo Clinic WSD dataset have more senses than the NLM WSD dataset. In all cases, the algorithm performs at or above the baseline of the majority sense.

5. Discussion

In this study, we addressed a wide variety of ambiguities, a total of 83 unique terms, from the clinical and scholarly biomedical domains—41 are unique to the Mayo Clinic dataset, 33 are unique to the NLM WSD dataset and 9 overlap between the two datasets. Each ambiguity was represented by a relatively small dataset of 100 instances except two which had 1000 instances. Indeed, we excluded only 8 of the NLM WSD dataset terms and added ambiguities, representing the most common ambiguities in the Mayo Clinic corpus of clinical notes. Others, also using the NLM WSD set for their studies, excluded 28 terms [8] and 35 terms [9]. The combination of many ambiguities and small sense-tagged datasets for each presented a number of challenges.

Our main finding is that a combination of features is necessary to achieve satisfactory results. A single feature was never discriminating enough across the two domains we tested—biomedical literature and clinical notes. Moreover, the most productive set of features varies across domains and terms—a consequence of the lexicalized nature of the features. A specific feature set is required to achieve optimal accuracy in sense disambiguation for each ambiguity which is a conclusion consistent with previous studies [8]. Indeed, this is the approach we took in the WSD component implementation within the Mayo Clinic text processing system. However, this approach is not generalizable and scalable to very large numbers of ambiguities, e.g. the 7400 ambiguous terms within UMLS, as it is prohibitively expensive to determine the most productive feature set for each specific ambiguity across potentially many domains and create enough sense-tagged instances. An alternative is to work with a generic set of the best performing features which in our study are *stemming, stopwords and punctuation removed, tf*idf by stem or token, BOW* within a small window size (5 or 10 tokens) and *orientation*. These best performing features can be used as the generic feature set to build WSD classifiers for any multisense word. That approach, however, does not address the problem of creating enough sense-tagged instances necessary to train the classifier.

This work was done with very small datasets which pose a challenge for any learning algorithm. Our plans are to pursue alternative techniques such as semi-supervised methods where we intend to utilize a small initial training corpus to generate a bigger pool of learning examples. Such an algorithm is presented in [23].

WSD is an important component within an information retrieval system as it has the potential to increase the pre-

cision of the retrieved documents hence lead to substantial cost-savings and satisfaction for the end users. However, we have not conducted an evaluation of our information retrieval system with the WSD component as part of it. This is yet another future investigative goal.

6. Conclusion

We investigated WSD across two domains and 83 unique ambiguities by applying a variation of Huber's algorithm to run experiments over 28 feature sets. For all experiments the results are at or well above the majority sense baseline. The most productive features that distinguished possible senses, were *stemming, removal of punctuation and stoplist words, filtering by a modification of the tf*idf metric* and *orientation*. Window size tended to be small—5 or 10 tokens on both sides. We conclude that it is not a single feature but a combination of features that generates the best results. This combination of features is different for each ambiguous term. Indeed, the best models for each ambiguous term are implemented in the Mayo Clinic text processing production system.

Acknowledgments

We thank James Buntrock and Vinod Kaggal for programming help, encouragement and insightful comments. We are very appreciative of the meticulous work the annotators did in sense-labeling the Mayo WSD dataset, and would like to acknowledge their outstanding effort—Barbara Abbott, Debra Albrecht, Pauline Funk, and Donna Ihrke.

References

[1] Shatkay H, Feldman R. Mining the biomedical literature in the genomic era: overview. J Comput Biol 2003;10:821–55.
[2] Schuler G, Epstein J, Ohkawa H, Kans J. Entrez: molecular biology database and retrieval system. Methods Enzymol 1996;266:141–62.
[3] Krallinger M, Valencia A. Text mining and information retrieval services for molecular biology. Genome Biol 2005;6:224.
[4] Aronson AR, Bodenreider O, Chang HF, et al. The NLM indexing initiative. Proc AMIA Symp 2000.
[5] Friedman C. Towards a comprehensive medical language processing system: methods and issues. Proc AMIA Symp 1997.
[6] Humphreys B, Lindberg D, Schoolman H, Barnett G. The unified medical language system: an informatics research collaboration. J Am Med Assoc 1998;5:1–11.
[7] Weeber M, Mork JG, Aronson AR. Developing a test collection for biomedical word sense disambiguation. Proc AMIA Symp 2001.
[8] Liu H, Teller V, Friedman C. A multi-aspect comparison study of supervised word sense disambiguation. JAMIA 2004;11(4).
[9] Leroy G, Rindflesch T. Using symbolic knowledge in the UMLS to disambiguate words in small datasets with a Naïve Bayes classifier. Proc Medinfo 2004.
[10] Xu H, Markatou M, Dimova R, Liu H, Friedman C. Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues. BMC Bioinformatics 2006;7:334.
[11] Sehgal AK, Srinivasan P, Bodenreider O. Gene terms and English words: an ambiguous mix. SIGIR'04 Workshop on Search and Discovery in Bioinformatics 2004.
[12] Agirre E, Edmonds P, editors. Word sense disambiguation: algorithms and applications. Springer; 2006.
[13] Schutze H. Automatic word sense discrimination. Comput Linguist 1998;24:97–123.
[14] Witten Ian H, Eibe Frank. Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann; 2005.
[15] Schuemie MJ, Kors JA, Mons B. Word sense disambiguation in the biomedical domain: an overview. J Comput Biol 2005;12:554–65.
[16] Leroy G, Rindflesch T. Effects of information and machine learning algorithms on word sense disambiguation with small datasets. Int J Med Inform 2005;74:573–85.
[17] Humphrey SM, Rogers WJ, Kilicoglu H, Demner-Fushman D, Rindflesch TC. Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: preliminary experiment. J Am Soc Inform Sci Technol 2006;57:96–113.
[18] Pakhomov SV, Pedersen T, Chute CG. Abbreviation and acronym disambiguation in clinical discourse. Proc AMIA Symp 2005:589–93.
[19] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.
[20] Carletta J. Assessing agreement on classification tasks: the kappa statistic. Comput Linguist 1996;22(2):249–54.
[21] Poesio M, Vieira R. A corpus-based investigation of definite description use. Comput Linguist 1998;24(2):183–216.
[22] Chklovski T, Mihalcea R. Exploiting agreement and disagreement of human annotations for word sense disambiguation. RANLP 2003.
[23] Ando RK. Applying Alternating Structure Optimization to Word Sense Disambiguation. In: Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL), 2006.
[24] Lee, Yoong Keok, Ng, Hwee Tou. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In: Proceedings of EMNLP-2002, 2002.
[25] Zhang T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: ICML 2004, p. 919–26.
[26] Coden A, Pakhomov S, Ando R, Chute C. Domain-specific language models and lexicons for tagging. JBI 2005.
[27] Cohen P. Empirical methods for artificial intelligence. Cambridge, Massachusetts: The MIT Press; 1995.
[28] Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrica 1934;26:404–13.